Assignment 2: K-means clustering

1 Instructions

- Use python programming language for your implementation.
- Use appropriate approach if you find some attribute is missing in your data.
- Report must contain step-wise description of your implementation and analysis of results. Since data analysis is a crucial task for any machine learning algorithm, report should demonstrate detailed analysis of results and conclusion. It should also clearly mention the steps to run your code.
- Do not use any direct in-built functions of libraries to implement K-means clustering algorithm, otherwise -10 will be deducted.

2 Dataset

- Pima Indians Diabetes Database: https://www.kaggle.com/uciml/pima-indians-diabetes-database. Download the dataset from the given link which also has description about the dataset.

3 Problem statement: K-means clustering

1. Implement K-means clustering algorithm for a user given K. **15**
2. Provide clustering performance i) using available ground truth, ii) without the ground truth. [metrics like homogeneity, ARI, NMI, etc] **5+5**
3. Provide the most suitable K for your data by varying K using some valid internal indices, (like silhoutte index, Calinski Harabasz, Wang's method of cross-validation.). Use graphical interpretation (K vs. metric) to justify your observation. **15**
4. Is there any change in clustering outcome if you initialize K cluster centroids with random points in different execution? (use the below given

Test-A for this purpose). If you observe any change, use any heuristic to initialize K centroids such that observed variation could be minimized? Use the heuristic to chose initial K centroids and perform Test-A. **15+10+15**

5. A brief report explaining the procedure and the results. **20**

Test-A

1. select K random points from the original data
2. split remaining data into 2 parts randomly (80:20) ratio
3. apply k-means on training data with K-random points selected in step-1
4. label test data using k-centroids.
5. use some metric(NMI,ARI, etc.) in test data to understand clustering accuracy
6. repeat 2-5 at least 50 times and report average metric

Repeat the above procedure for 50 different K-random points. Report the dispersion of the metric.